IBM BY COURSERA

Applied Data Science Capstome Project

OPENING A NEW SPORT UNIT IN BERLIN

By Vera Rykalina

January, 2020

Contents

1 Introduction

1.1 Berlin City

Berlin is the largest city in Germany with almost 4 million inhabitants. Not only is Berlin the capital of Germany 1 but also it is the second most populous city of the European Union. The city has its special vibrant atmosphere and full of multicultural communities what appeals foreigners from around the world. Due to its high quality of life, Berlin is truly considered to be a scientific hub attracting specialists from different industrial and academic sectors.

Figure 1: Berliner Ampelmann

However, the most valuable characteristics for living in Berlin is its ecological situation. Berlin is a green city. What doest it mean? First of all, there are dozens of parks, forest areas and recreation spots. Moreover, people separate rubish, recycle plastic, ride bycicles and are obsessed about bioshops in a good way. In addition to it, and what is especially inportant, people in Berlin strive to lead a healthy life style. The inhabitants prefer clean eating and actively do sports. In this report I would like to consider a possibility to open a new sport unit in Berlin preferably in the localities which are lacking sport centers. The bussiness

decision should be made based on several parameters, mainly: a location, a population of the district and a number of the neighbouring sport units.

1.2 Business Challenge

The objective of the given capstone project is to analyse and select the best locations in Berlin city to open a new sport unit (gym or fitness studio). Using data science methodology and machine learning algorithm such as clustering, the capstone project aims for providing solutions to answer the fowllowing business question: What are the best localities for a new sport unit in Berlin?

1.3 Target Audience

The given project can be particularly useful for a potential property developer or an investor who is planning to launch or invest in a new fitness unit in Berlin. The project goal is very challenging as the city is currently having a great number of sport spots but on the other hand, very fast growing dynamic and active population of Berlin is demanding in terms of variety of sport activities and vacant memberships in already existing sport centers. Besides, there are localities, which have a limited number of the sport schools and centers due to historical or economical reasons but still with increasing demand for such units.

2 Data

2.1 Type of Data

To answer the business question, one should collect the following data:

- A list of the localities in Berlin.
- Demographic characteristics of each locality.
- Latitude and longitude coordinates of the localities. This type of data is required for locality mapping and venue data establishment.
- A venue data preferentially anchored in the data related to a gym or a fitness center for clustering.

2.2 Data Mining

Berlin city consists of 95 localities. A list of the localities of the German capital can be found using a Wikipedia page https://en.wikipedia.org/wiki/Category:Localities_of_Berlin. The Web scraping for the data is performed using two Python packages: Requests and Beautifulsoup. The geographical coordinates of the localities are obtained by Python Geocoder package. Foursquare API is leveraged for the venue data acquisition. The Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API can provide a developer with many categories of the venue data. The focus of this projects is a category linked with fitness centers.

To sum up, the given project should make use of many data science skills from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and geovisualization (Folium).

3 Methodology

The list of localities of Berlin was obtained from the open source Wikipedea page using web scapping Python libraries (Requests and Beautifulsoup packages). Geocoder package was utilized for identification of the geographical coordinates for Berlin localities (latitude and longitude parameters). Collection of data corresponding to localities with geographiclal coordinates was tranformed to a pandas DataFrame and then visualized as a map using Python Folium package. The establishment of top 100 venues within a radius of 2000 meters was performed by means of Foursquare API. Foursquare returns the venue data in JSON format with the possibility to extract a venue name, a venue category and venue latitude and longitude. Unique venue categories were systemized and counted using Python pandas methods. As the last data preroccessing step the venue rows were groupped by the mean of the frequency of occurrence of each venue category and filtered by gym/fitness studio category. Being groupped and filtered the data was used for analysis with K-means algorithm. The algorithms identified k number of centroids and then allocated every data point to the nearest cluster, while keeping the centroids as small as possible. Being a simple unsupervised machine learning method, K-means is a good fit for the project task. 3 clusters were determined based on the frequency of occurrence for gym/fitness studio category.

4 Results

The result from the k-means clustering demonstrates clear distrubution of clusters based on different frequency of occurrence of the gym/fitness studios 2:

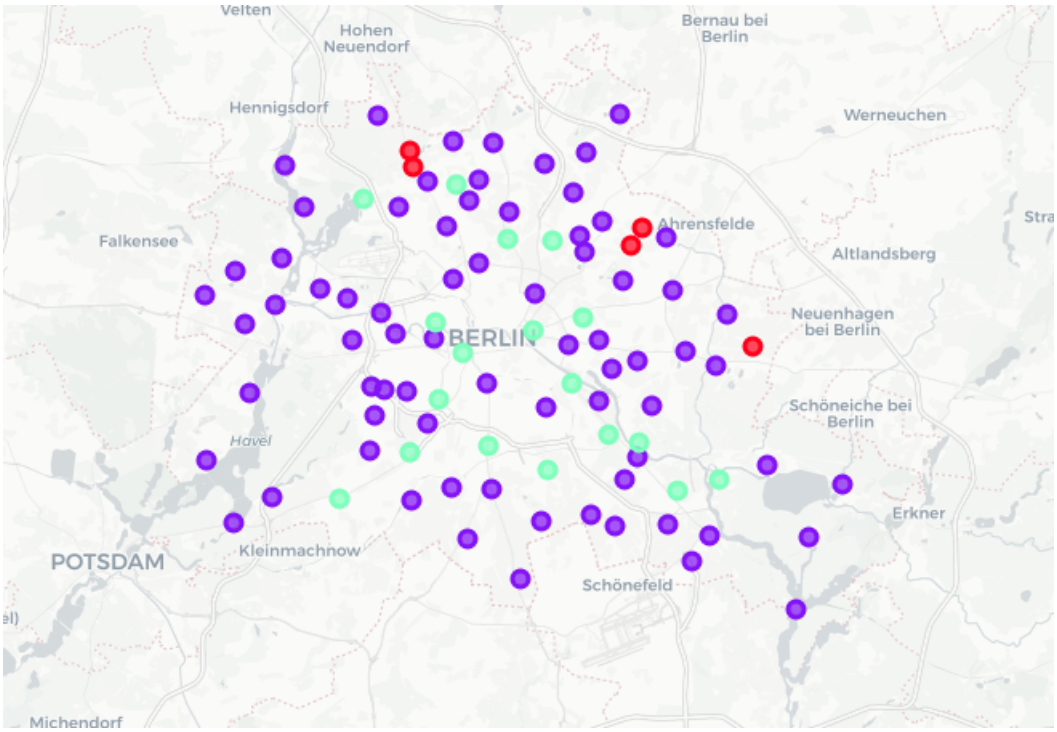

Figure 2: The result of the clustering is visualized by the map with cluster 0 in red, cluster 1 in purple and cluster 2 in mint colour

- Cluster 2: Localities with moderate number of fitness centers
- Cluster 1: Localities with low number to no existence of fitness centers
- Cluster 0: Localities with high concentration of fitness centers

5 Discussion

Using Python pandas filters top 20 localities were selected and analyzided according to density inhabitants per square km and area in square km. The results are visualizid using Seaborn library and presented on Figure 3 and Figure 4. As can be seen Friedenau, Fennpfuhl, Kreuzberg and Gesundbrunnen are the most

densely populated although for instance Friedenau and Fennpfuhl are not the largest districts. Districts covering the area more than 20 square km are Köpenick, Tegel, Wannsee and Grunewald.

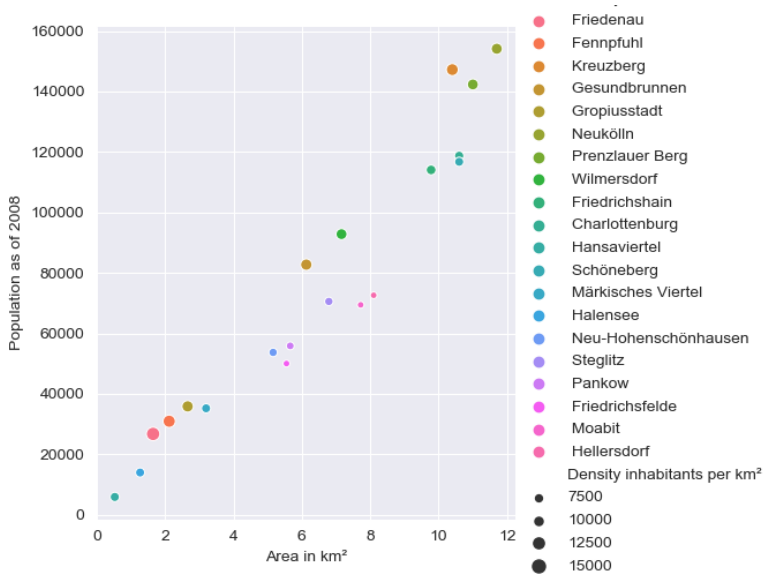

Figure 3: A scatter plot illustrating a relationship of population and loclity area

As for the cluster analysis, most of the gyms or fitness centers with moderate concentration are located in the city center.The highly concentrated sport spots (cluster 0) takes place in the north-eat part of Berlin.These sport units were formed due to geographical convenience and historical reasons. On the other hand, cluster 1 has a very low number to no fitness centers in the locality what can be interpreted as a great opportunity for launching a new sport unit. Besides, sport unint in these localitues are far away from the city center and could be a reasonable investment.

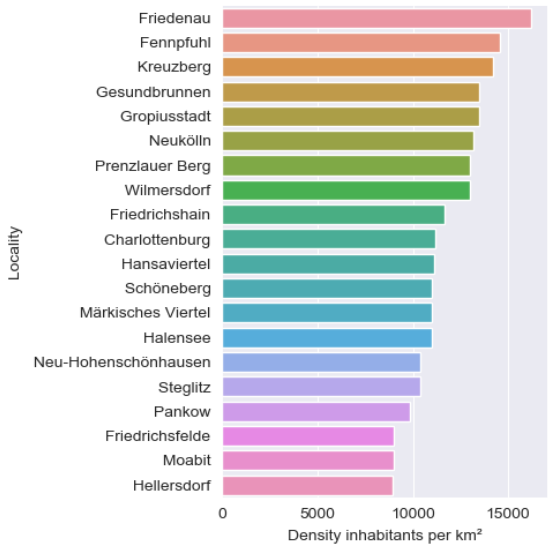

Figure 4: Population on sorted top 20 localities

It is also worth mentioning the proving the data against the relvant category. Using Pandas the final dataframe was sorted to indentify top 15 categories of venues. Data visualysation results using Python Seaborn lbrary are shown in Figure 5. It is clearly seen the our target fitness studio category is not within the ranking which can be a good indication that this category is underrepresented.

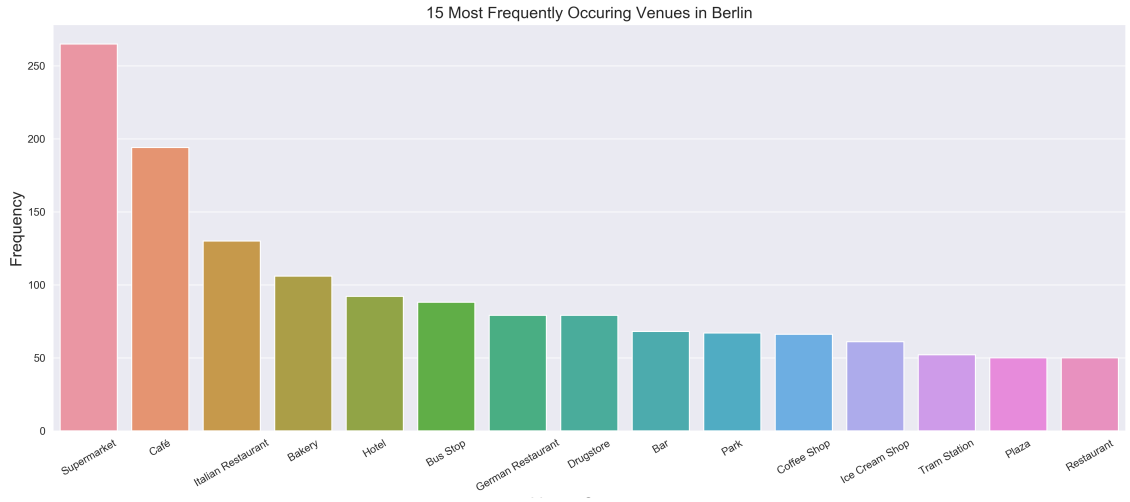

Figure 5: Top 15 venue categories

6 Recommendations

Capstone project recommendations to property developers to make thier investments based on the clustering findings:

- Investors are advised to consider launching a new sport unit in any locality in cluster 1 with low competition level and lower building price per square meter than in the city center.
- Stakeholders are also encouraged to look for opportunities to open new sport unit in any locality in cluster 0, if they are ready for a moderate competition conditions.
- Potential property developers should avoid localities in cluster 2 which already have high concentration of gyms and fitness studios.

7 Conclusions

In the given project several steps were taken for stating a business problem, specifying the requried data, extracting and preparing the data and performing machine learning clustering algorithm to group the localities into 3 clusters based on their similarities. Moreover, the useful recommendations were provided to the potential stakeholders i.e. property developers and investors regarding the best locations to open a new fitness center. Back to the business question framed in the introduction section, the localities in cluster 1 are the most preferred locations to open a new gym or a fitness center. The findings of this project can be helpful to relevant stakeholders to make their investments in the reasonable way and get payoffs in short period of time.

References

[1] Localities in Berlin. Wikipedia.

<https://en.wikipedia.org/wiki/Category:Localities_of_Berlin>

[2] Boroughs and neighborhoods of Berlin. Wikipedia.

<https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin>

[3] Foursquare Developers Documentation. Foursquare.

<https://developer.foursquare.com/docs>